Rancang Bangun Sistem Pengelolaan Dokumen-dokumen Penting Menggunakan Text Mining

Ahmad Hatta A [1)], Nana Ramadijanti, S.Kom, M.Kom [2)], Afrida Helen, S.T., M.Kom [2)]
Mahasiswa[1] , Dosen [2]
Jurusan Teknik Informatika
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember
Kampus PENS-ITS Keputih Sukolilo Surabaya 60111
Telp (+62)31-5947280, 5946114, Fax. (+62)31-5946114
E-mail : aahmadhatta@gmail.com , nana@eepis-its.edu, helen@eepis-its.edu

Abstrak

Saat ini kebutuhan akan penggunaan data dalam format digital semakin meningkat. Terbukti dengan semakin berkembangnya teknologi smart card yang digunakan untuk menggantikan data-data penting ke dalam satu kartu. Kebutuhan ini semakin terasa ketika berhubungan dengan penyimpanan arsip-arsip yang penting. Hal ini dikarenakan format digital dirasakan lebih efektif dan efisien. Namun proses untuk mengubah data dari format kertas menjadi data format digital memerlukan usaha yang sangat besar, baik dari tenaga, waktu maupun biaya yang dibutuhkan. Maka untuk memperkecil usaha yang dibutuhkan perlu dibuat suatu sistem pengelolan dokumen digital dengan menggunakan teknologi OCR (*optical character recognizing*) guna merubah data menjadi data digital serta melakukan pengelolaan dokumen digital tersebut berdasarkan informasi yang terkandung didalamnya. Untuk memudahkan sistem untuk mendapatkan informasi didalam dokumen tersebut dapat dilakukan dengan menerapkan tahapan-tahapan *text mining*. Sedangkan pengelolaan dokumen menggunakan metode *K-Nearest Neighbors* (K-NN) untuk mengelompokkan data digital tersebut sesuai dengan tipe masing-masing.
Kata kunci : optical character recognizing, nearest neighbors, text mining, dokumen digital.

1. Pendahuluan

Dari dahulu kebanyakan dokumen dalam bentuk kertas yang didalamnya terkandung informasi yang bisa dibaca oleh manusia secara umum. Kertas merupakan media yang sering digunakan untuk menampilkan isi dokumen dan dalam bentuk inilah suatu dokumen disimpan dan dikelola. Namun sering kali seseorang merasa kesulitan jika harus membawa dan mengelola dokumen dalam jumlah besar.

Seiring perkembangan jaman, media kertas mulai sedikit demi sedikit ditinggalkan, dikarenakan kertas dirasa kurang efektif dan efisien. Sekarang seseorang lebih memilih untuk membawa data dalam format digital dari pada kertas. Sehingga dari waktu ke waktu, penggunaan data digital semakin meningkat apalagi ketika berhubungan dengan penyimpanan arsip-arsip yang penting. Format digital dirasakan lebih efektif dan efisien, karena data lebih mudah disimpan dan mudah dicari ketika dibutuhkan.

Proses perpindahan data dari format kertas menjadi data dengan format digital memerlukan usaha yang sangat besar, baik dari segi tenaga yang dibutuhkan , waktu yang diperlukan maupun dari segi biaya yang dibutuhkan. Namun apabila tidak dilakukan perubahan data tersebut, maka usaha yang dibutuhkan untuk melakukan pemeliharaan , pencarian, penggunaan , serta penjagaan akan jauh lebih besar.

Maka konsekuensinya adalah penyediaan teknologi yang dapat mengubah data dari format kertas ke dalam digital kemudian mengeluarkan / menyimpan informasi yang terkandung dalam data dengan format digital tersebut.

Oleh karena itu, pembuatan tugas akhir ini diharapkan dapat membantu memberikan kemudahan dalam hal penyimpanan dokumen-dokumen penting kedalam format digital bagi masyarakat khususnya instansi-instansi pemerintahan yang masih menyimpan dokumen-dokumen pentingnya dalam format kertas.

2. Dasar Teori

2.1 *Optical Character Recognizing (OCR)*

OCR adalah sebuah sistem komputer yang dapat membaca huruf, yang berasal dari sebuah pencetak (*printer* atau mesin ketik). Adanya sistem pengenal huruf ini akan

meningkatkan fleksibilitas ataupun kemampuan dan kecerdasan sistem komputer. Dengan adanya sistem OCR maka *user* dapat lebih leluasa memasukkan data karena *user* tidak harus memakai papan ketik tetapi bisa menggunakan pena elektronik untuk menulis sebagaimana *user* menulis di kertas. Adanya sistem pengenal huruf yang cerdas akan sangat membantu usaha besar-besaran yang saat ini dilakukan banyak pihak yakni usaha digitalisasi informasi dan pengetahuan, misalnya dalam pembuatan koleksi pustaka digital, koleksi sastra kuno digital, dan lain-lain.

2.2 *Text Mining*

Text Mining adalah suatu proses yang bertujuan untuk menemukan informasi atau tren terbaru yang sebelumnya tidak terungkap, dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu [2]. Selain itu *text mining* juga diartikan sebagai kegiatan menambang data dari data yang berupa teks atau dokumen, dengan tujuan mencari kata-kata yang dapat mewakili apa yang ada dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Langkah-langkah yang dilakukan dalam *text mining* adalah sebagai berikut:

- *Tokenizing*

Proses ini memotong setiap kata dalam teks, dan mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai 'z' yang diterima, sedangkan karakter selain huruf dihilangkan. Jadi hasil dari proses *tokenizing* adalah kata kata yang merupakan penyusun kalimat/*string* yang dimasukkan.

- *Filtering*

Pada tahap ini dilakukan proses filter atau penyaringan kata hasil dari proses *tokenizing*, dimana kata yang tidak relevan dibuang. Proses ini menggunakan pendekatan *stoplist*. Yang termasuk *stoplist* adalah "yang", "di", "dari", dan lain-lain.

- *Stemming*

Stemming adalah proses untuk menggabungkan atau memecahkan setiap varian-varian suatu kata menjadi kata dasar. *Stem* (akar kata) adalah bagian dari akar yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran).

- *Tagging*

Tagging adalah suatu proses mencari bentuk asal dari kata bentuk lampau. Tahap ini tidak digunakan pada teks berbahasa indonesia karena kata dalam bahasa indonesia tidak mempunyai bentuk lampau.

- Analizing

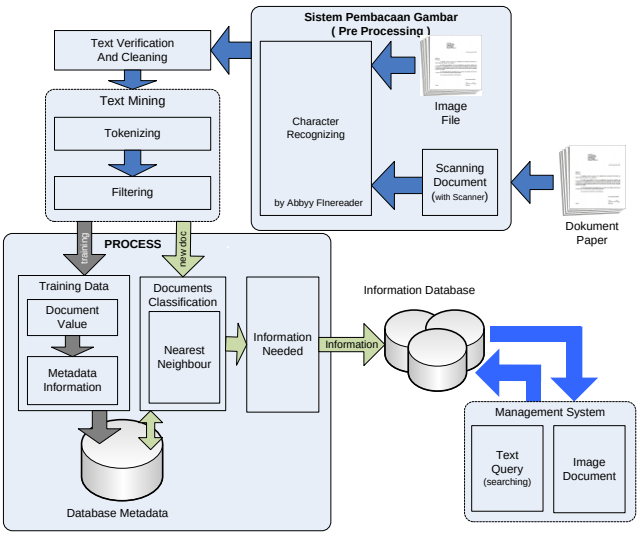
Pada tahap ini dilakukan proses perhitungan bobot (w) dokumen agar diketahui seberapa jauh tingkat similaritas antara *keyword* yang dimasukkan dengan dokumen.

2.3 *Nearest Neighbors*

Nearest Neighbors merupakan suatu metode untuk mengklasifikasikan suatu data baru berdasarkan similaritas dengan label data. K-*Nearest Neighbor* merupakan salah satu metode yang digunakan dalam pengklasifikasian. Prinsip kerja K-*Nearest Neighbor* (KNN) adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (*neighbor*) terdekatnya dalam data pelatihan. Algoritma k-NN adalah sebagai berikut :

- Tentukan k.
- Hitung jarak antara data baru ke setiap labeled data.
- Tentukan k labeled data yang mempunyai jarak yang paling minimal.
- Klasifikasikan data baru ke dalam *labeled* data yang mayoritas.

3. Perencanaan Sistem



Gambar 1 *blok diagram umum sistem*

3.1 *Preprocessing*

Pre-processing adalah proses awal mengelola data sebelum pengolahan data yang dilakukan. Proses ini bertujuan untuk merubah data dalam format kertas ke format digital dengan melakukan proses *scanning* terhadap data yang dimiliki dengan menggunakan *scanner* dan kemudian melakukan pembacaan gambar sehingga didapat informasi berupa teks.

3.1.1 Scanning

Proses scanning dilakukan dengan memakai alat bantu berupa scanner dan perangkat lunak pendukung

yang berguna untuk mengambil data gambar hasil pembacaan scanner. Untuk kedalaman gambar pembacaan gunakan nilai *dpi (dot per inch)* minimal 300 dpi.

3.1.2 Pembacaan Karakter

Proses pembacaan karakter dari gambar yang telah disimpan dari hasil proses editing gambar. Proses ini bertujuan untuk mendapatkan karakter-karakter/informasi yang terdapat didalam gambar tersebut. Untuk melakukan pembacaan Karakter ini menggunakan OCR (*optical character recognizing)*.

3.2 Pengolahan Hasil Pembacaan Karakter menggunakan Text Mining

Proses pengolahan hasil pembacaan karakter dilakukan dengan menggunakan beberapa tahapan *text mining* diantaranya *tokenizing* dan *filtering*.

Penjelasan singkat mengenai proses *text mining* yang dilakukan :

1. *Tokenizing*

Didalam proses *tokenizing* yang dilakukan yaitu memecah kalimat hasil pembacaan karakter menjadi kumpulan kata-kata.

2. *Filtering*

Proses *filtering* dilakukan setelah didapatkan kumpulan kata-kata kemudian dilakukan pemilihan kata-kata penting atau dilakukan pembuangan kata-kata yang dianggap tidak perlu

3.3 *Document Training*

Pada proses *document training* ini akan dilakukan pembelajaran ciri-ciri yang dimiliki oleh suatu dokumen tertentu. Hasil proses pembelajaran tersebut kita ujicoba dengan menggunakan data test untuk memberikan label dokumen tersebur (*test data*).

Untuk memberikan label pada data tes sesuai dengan jenisnya dilakukan perhitungan jarak (*euclidean*) nilai metadata data tes dengan nilai metadata hasil pembelajaran ciri khusus masing-masing dokumen.

Tabel 4. 1 *ilustrasi data training*

Keyword	d1	d2	dn
No.KTP	1	0	0
Nama	1	1	1
Tempat Lahir	1	0	0
Tgl Lahir	1	1	0
Alamat	1	0	0
J.Kelamin	1	0	0
Pekerjaan	1	0	0
……...	...	…	…
Pelaksana	0	0	0

KTP	SK	KPG
1	1	1
1	0	1
1	0	1
1	0	0
1	0	0
1	1	0
1	1	0
…	…	….
1	0	0

Proses *document clustering* ini dilakukan satu kali ketika diawal, yang digunakan untuk membentuk *database metadata*.

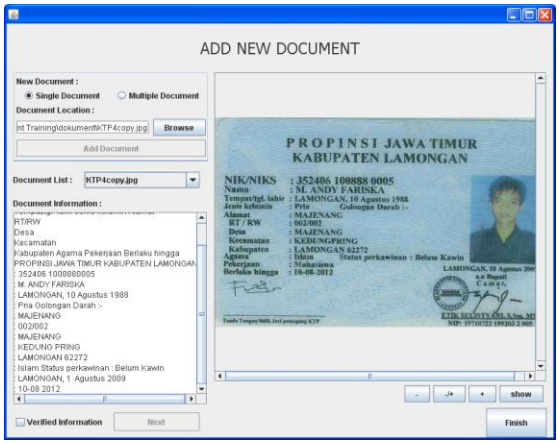
3.4 *Document Classification.*

Pada proses *document classification* ini akan dilakukan proses klasifikasi dokumen-dokumen baru yang akan ditambahkan ke dalam database. Untuk mengklasifikasikan dokumen baru ini kita akan memerlukan hasil proses *clustering* yang telah didapat pada proses *document clustering* diatas. Proses yang dilakukan untuk mengklasifikasikan data baru tersebut termasuk ke dalam jenis data tertentu dapat kita lakukan dengan menggunakan algoritma NN (*Nearest Neighbour*) atau lebih tepatnya k-NN dengan semakin besar nilai k maka akan semakin besar nilai ketepatannya.

4. Hasil dan Pembahasan

Setelah dilakukan pengujian terhadap system yang dibuat, diketahui bahwa sistem dapat berjalan lancar apabila data teks hasil dari pembacaan gambar dalam keadaan lengkap dan sesuai dengan data yang terdapat pada dokumen asli.

4.1 Analisa proses penambahan dokumen baru



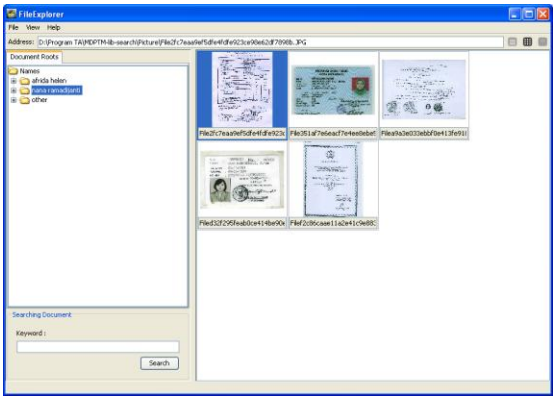
Pada hasil pembacaan dokumen baru diatas didapatkan hasil pembacaan yang kurang bagus dan akurat pada beberapa dokumen terlihat dengan banyaknya kata atau karakter yang hilang serta munculnya karakter-karakter aneh.

No	Percobaan	Kata yang hilang / Error	Karakter Aneh
1	Percobaan 1 (KTP)	Error: WAUKOTA,PJLKEPALA, KEPENOUDUKAN,SIPSL Hilang : Pekerjaan	-
2	Percobaan 2 (KTP)	Error: 2M1-1965, NGINDENJANGKUNGAN, Kaviin,LA,TangaiVSidik, NGINDEN1NTAN Hilang : NIK,Nama,Tempat/Tgl Lahir,Jenis kelamin, Alamat,Agama,Pekerjaan,Berlaku Hingga, Kewarganegaraan	-
3	Percobaan 1 (Ijazah)	Wiratmo p Viiwono dim mduh satu 9fopem6erseriSu semSilan ratus tujun piifufi sembuan Te ibJEkfyro - Te ni Sistem Xpmputer	-

Analisa penyebab jeleknya hasil pembacaan :

1. Keadaan dokumen fisik yang kurang baik
2. Terdapat banyak *noise* pada data seperti *border*, bercak, bekas stempel dan sebagainya.
3. Hasil pembacaan OCR yang kurang baik.

Sistem penambahan data baru dapat berjalan dengan lancar. Dari 5 kali melakukan uji coba penambahan dokumen baru, sistem dapat berjalan sesuai dengan keinginan apabila data/informasi yang terkandung didalam data baru dapat terbaca dengan baik dan lengkap.



Setelah penambahan data baru berhasil, hasil penambahan ditampilkan ke dalam program explorer diatas dikelompokkan ke dalam pemilik dan jenis dokumen tersebut.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil pengujian sistem yang telah dilakukan didapatkan beberapa kesimpulan :

- Hasil pembacaan dokumen menggunakan OCR sangat berpengaruh terhadap sistem,
- Proses pengenalan ciri-ciri dokumen dipengaruhi oleh banyaknya keanekaragaman dokumen.
- Proses klasifikasi data baru dengan K-NN memberikan hasil yang akurat kecuali dokumen Sertifikat. Karena sertifikat yang beredar memiliki variasi yang bermacam-macam.

5.2. Saran

Hasil dari proyek akhir ini belumlah sempurna, untuk meningkatkan hasil yang dicapai dapat dilakukan:

- Penggunaan Software OCR yang lebih baru/lebih canggih sehingga didapatkan hasil pembacaan yang lebih baik.
- Perlunya dikembangkan sistem untuk perbaikan hasil pembacaan secara otomatis.

6. Daftar Pustaka

[1] Hermaduanti, Ninki., Kusumadewi,Sri. 2008.*Sistem Pendukung Keputusan Berbasis Sms Untuk Menentukan Status Gizi Dengan Metode K-Nearest Neighbor*. Yogyakarta: Universitas Islam Indonesia.

[2] Kunaifi,Aang.2009.*Klasifikasi Email Berbahasa Indonesia Menggunakan Text Mining Dan Algoritma KMeans*. Surabaya: Politeknik Elektronika Negeri Surabaya.

[3] Trilaksono,Mirza. 2008. *Implementasi Optical Character Recognition (Ocr) Dengan Pendekatan Metode Struktur Menggunakan Ekstraksi Ciri Vektor Dan Region IT Telkom*.From http://www.ittelkom.ac.id/library/index.php?view=article&catid=11:sistem-komunikasi&id=93:ocr-optical-character-recognition&option=com_content&Itemid=15 (akses :Januari 2010).

[4] Fan, Weiguo.(2006). *Text Mining, Web Mining, Information Retrieval and Extraction from the WWW References*. From http://filebox.vt.edu/users/wfan/text_mining.html. (Akses : Februari 2010)